Acoustic Species Identification
Milestone Report

Ludwig von Schoenfeldt **Sean O’Brien** **Vibhuti Rajpurohit** **Geelon So**

1 Minimum Viable Product

1.1 Project Overview

Gathering and analyzing data on the prevalence of bird populations in a given region is key to ecological research. However, the pace of data collection has far outstripped the ability of researchers to manually analyze and label it.

In particular, we focus on the task of identifying bird calls present in acoustic recordings from soundscapes. Previous work from the Engineers for Exploration team has established PyHa, a library for identifying potential avian vocalizations, and more recently a classification pipeline to automate data augmentations, model training and evaluation.

Our project builds on this work in three core research directions: **data representation**, **ensembling strategy** and **model architecture**, with the goal of improving pipeline accuracy as measured by performance on the BirdCLEF 2024 contest.

1.2 MVP Description

Our MVP integrates a suite of scripts into the existing acoustic multi class training pipeline, facilitating the training and deployment of an ensemble of machine learning models for enhanced classification of bird vocalizations. This enhancement is principally realized through the introduction of two primary scripts:

1. A script has been introduced to enable sequential training utilizing various machine learning models within the established multi class training pipeline.

2. An ensemble script has been implemented to aggregate the outputs from multiple trained machine learning models. This script employs averaging techniques across the predictions of individual models to optimize the overall prediction accuracy.

2 Group Management

Ludwig is the project lead of the team and has been overseen tasks and supported the team by evaluating progress in order to maintain the outlined plan. Our general decisions so far have been made through consensus. We also received advice and input from the other project leads from E4E, namely Sean Perry and Sam Prestrelski and have continuously updated and discussed progress with the collaborators from the San Diego Zoo Alliance during the E4E Acoustic Species ID weekly collaborator meetings. We have communicated via Discord to stay updated and share results and adjust our plan. At the same time, we also meet up during the work sessions of the E4E Acoustic Species ID team to discuss progress and refine plans for the following week. Besides that we have been meeting during some of the weeks for smaller updates or assignments for the course.

Ludwig has consistently updated both the E4E team at the general meeting and the San Diego Zoo collaborators at the collaborator meeting together with Geelon and Vibhuti about our current progress. The team also has had smaller work sessions in pairs, for example for pair programming.

2.1 Development Roles

Our project can mostly be divided into 3 separate categories: Research, Development and Evaluation. During the initial stage of this project we mostly conducted research into several model architectures and approaches to the problem with a special focus on newer non computer vision approaches. Towards the current second half of the project we are focusing on evaluation of the newer models comparing it to already existing computer vision based approaches to create a fully functional architecture which we will then submit on Kaggle for BirdCLEF 2024. The role assignment:

Ludwig: Lead, Research (Ensembling), Evaluation, Development

Geelon: Research (Representation), Development, Evaluation

Sean: Research (Architecture), Development, Evaluation

Vibhuti: Development, Evaluation

3 Project Progress and Milestone Completion

3.1 Architecture

Most approaches to classifying bird vocalizations are based on convolutional neural networks (CNNs) operating over a spectrogram generated from the raw audio.

However, recent literature has shown the merits of a new class of recurrent models, called Structured State Space Models, on several long-context sequence modeling tasks like speech modeling and DNA classification.

The E4E team has briefly experimented with applying such models to the BirdCLEF task with limited success. However, since that exploration an improved sequence modeling architecture, Mamba, has shown even further improvement.

The first step was implementing and validating Mamba-S4, a mixer model described in the Mamba paper but not publicly implemented. Next, the model was integrated into the classification pipeline. Finally, we conducted a short hyperparameter search over training parameters to train the Mamba-S4 model effectively.

The current effort is to determine the number of Mamba-S4 layers, effective sampling rate, batch size and learning rate to obtain stable training dynamics for the network. The status of this arm of the project is summarized in the table below:

Status	Milestone
Completed	Implement Mamba-S4
Completed	Incorporate Mamba-S4 into classification pipeline
In progress	Conduct hyperparameter search for training
In progress	Evaluate Mamba-S4 on validation set
To Do	Incorporate Mamba-S4 into ensemble pipeline
To Do	Run Mamba-S4 classification on distilled time series representation

3.2 Representation

Audio data has two natural representations, one in the *time domain* and another in the *frequency domain*. In the time domain, sound is directly represented by its acoustic wave. One difficulty of working in the time domain is the long-term dependencies, where the length of the context increases with the sampling rate, easily reaching tens of thousands of samples per second.

Classical methods therefore tend to work in the frequency domain, where the time series is decomposed into its component frequencies by the Fourier transform. The spectrogram is obtained by applying the Fourier transform on a sliding window over the sound wave—it measures how the contribution of each component frequency varies over time. We can view this as a simple and mathematically-sound feature extraction method.

While this representation has mathematically nice properties, it is *oblivious* to the data and therefore does not necessarily align with downstream tasks, nor with human perception. For instance, this latter

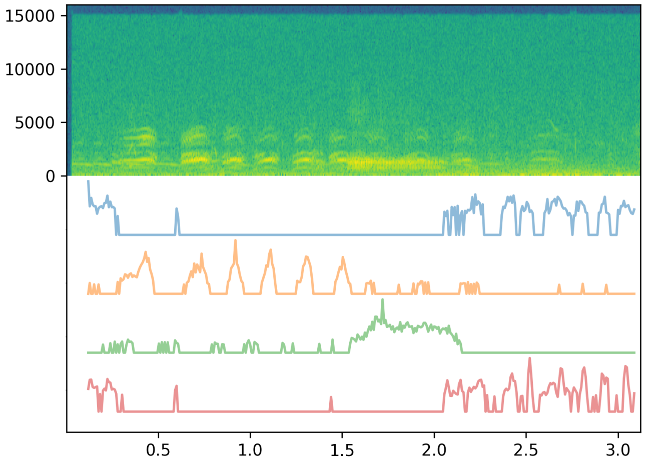


Figure 1: An example of a data-dependent representation under a four-element basis.

misalignment motivates the mel spectrogram, which logarithmically rescales the frequencies so as to more closely match how humans perceive differences in frequency.

This arm of the project seeks to understand whether there is any benefit to using a *data-dependent* representation of sound. Instead of decomposing a sound wave using the Fourier basis, we aim to construct a new basis that is learned in an unsupervised manner using a large set of bird call clips.

The first step was to implement a module that computes a new representation of a sound wave given any choice of basis. For example, when given the Fourier basis, it returns the standard spectrogram. Additionally, we can also compute common non-linear transformations/filters. In particular, top-k/top-p-quantile thresholding and binary quanitization.

Next, we implemented a simple heuristic for extracting a basis elements from a large set of bird call clips. This is achieved by randomly sampling a set of reference clips, and extracting short sequences that are self-predictive (more formally, a sequence that has maximum cosine similarity with the rest of the clip). Figure 1 visualizes the representation of a sound clip using four basis elements extracted in this manner. In the accompanying files, we provide audio of a collection of basis elements learned in this way.

A key remaining task is to incorporate the new representations into the Mamba-S4 implementation and existing CNN-based models. If time permits, we would also like to experiment with other standard methods for extracting prototypes for the new basis, including k-means clustering and dictionary learning/sparse coding. However, as per the feedback, this may be too ambitious for the time remaining, and so we have shifted to incorporating this MVP version of the representation into the downstream models.

Status	**Milestone**
Completed	Implement representation module
Completed	Implement heuristic feature extractor
In progress	Incorporate distilled time series representation into models
Stretch	Experiment with additional feature extraction methods

3.3 Ensembling Strategy

Reviewing past BirdCLEF competitions, it has become increasingly apparent that a technique called "Ensemble Learning" where several pretrained machine learning models (mostly CNNs, eg. EfficientNet, ResNET, etc.) are being used together in an ensemble to increase the accuracy in predicting bird species through their vocalizations.

To implement this, the team trained 7 different CNNs based on 5 major architectures, namely: EfficientNet, ResNET, Seresnet, RexNET and Convnext. Figure 2 shows the training loss plots of these models. In the meantime we have also started to look into several approaches for evaluating the ensemble ranging from averaging the models, assigning different weights to models, picking models based on the maximum results for a particular species to assigning each model to a specific species for prediction and evaluation.

We also wrote a script in the config file within the acoustic-multiclass-training pipeline to sequentially train several models within the given pipeline which enables an easy way to directly edit for each sequential run the model that should be used, the epochs and the learning rate for that specific run.

As part of the MVP, we have submitted a baseline model to the Kaggle competition. Figure 3 shows a screenshot of the successful submission.

Status	Milestone
Completed	Trained 7 different CNN models with BirdCLEF 24 data
Completed	Script to run several ML models sequentially within the existing training pipeline
Completed	Created notebook on Kaggle to evaluate models
In progress	Evaluate results from every model
In progress	Writing a script to evaluate ensemble through averaging method
To Do	Explore and try other evaluation methods for the ensemble
To Do	Potentially incorporate Mamba-S4 classification into ensemble

4 Remaining Schedule and Timeline

Two important feedback we have received is that (1) we need to ensure that we have a realistic timeline and target deliverable, and (2) that we have not been able to ‘fail fast’. Therefore, we have updated our main goal from research to implementation, to better evaluate the progress of the research phase. In particular, (a) in the Architecture arm, the focus is on hyperparameter search, (b) in the Representation arm, the focus has shifted to integration with existing models, and (c) in the Ensemble/Kaggle submission arm, the goal is to iteratively improve on the baseline we have provided as part of the MVP.

4.1 Deliverable Architecture

Milestone	Time Estimate	Teams
Conduct hyperparameter search for training	3 Days	Development
Evaluate Mamba-S4 on validation set	3 Days	Development

4.2 Deliverable Representation

Milestone	Time Estimate	Teams
Incorporate distilled time series representation into models	7 Days	Development

4.3 Deliverable Ensembling Strategy

Milestone	Time Estimate	Teams
Evaluate results from every model	2 Days	Evaluation
Script to evaluate ensemble through averaging method	4 Days	Development
Explore and try other evaluation methods for the ensemble	8 Days	Research
Incorporate Mamba-S4 classification into ensemble	2 Days	Development

5 Feedback

5.1 Elevator Pitch Feedback

The initial pitch got pretty good feedback, with the audience really seeing the potential for how this project could help the biodiversity ecosystems. Some asked how the bird data actually helps ecosystems, pointing out that it'd be great to make the real-world benefits clearer right from the start. The project's introduction seemed to lack in grabbing the audience's attention more, maybe by highlighting exactly why it is so important. Some suggested to also emphasise more on why this project is important and the need to be implemented immediately.

5.2 Project Specification Feedback

Feedback on project planning emphasized the need to set up a clear decision-making process before any issues pop up as making decisions early on can prevent confusion and delays later. We were also suggested to focus on building a proactive framework for making choices as a team. This means laying out how decisions will be made, who decides what, and what to do if people disagree. Other points suggested improving a few technical details in our approach.

In the technical side, it was mentioned that the project description could be misleading, especially around what the current tools can do. It's crucial that the technical aspects, like the use of PyHa models, are described accurately to avoid confusion. Better explanations here can help everyone understand the project better.

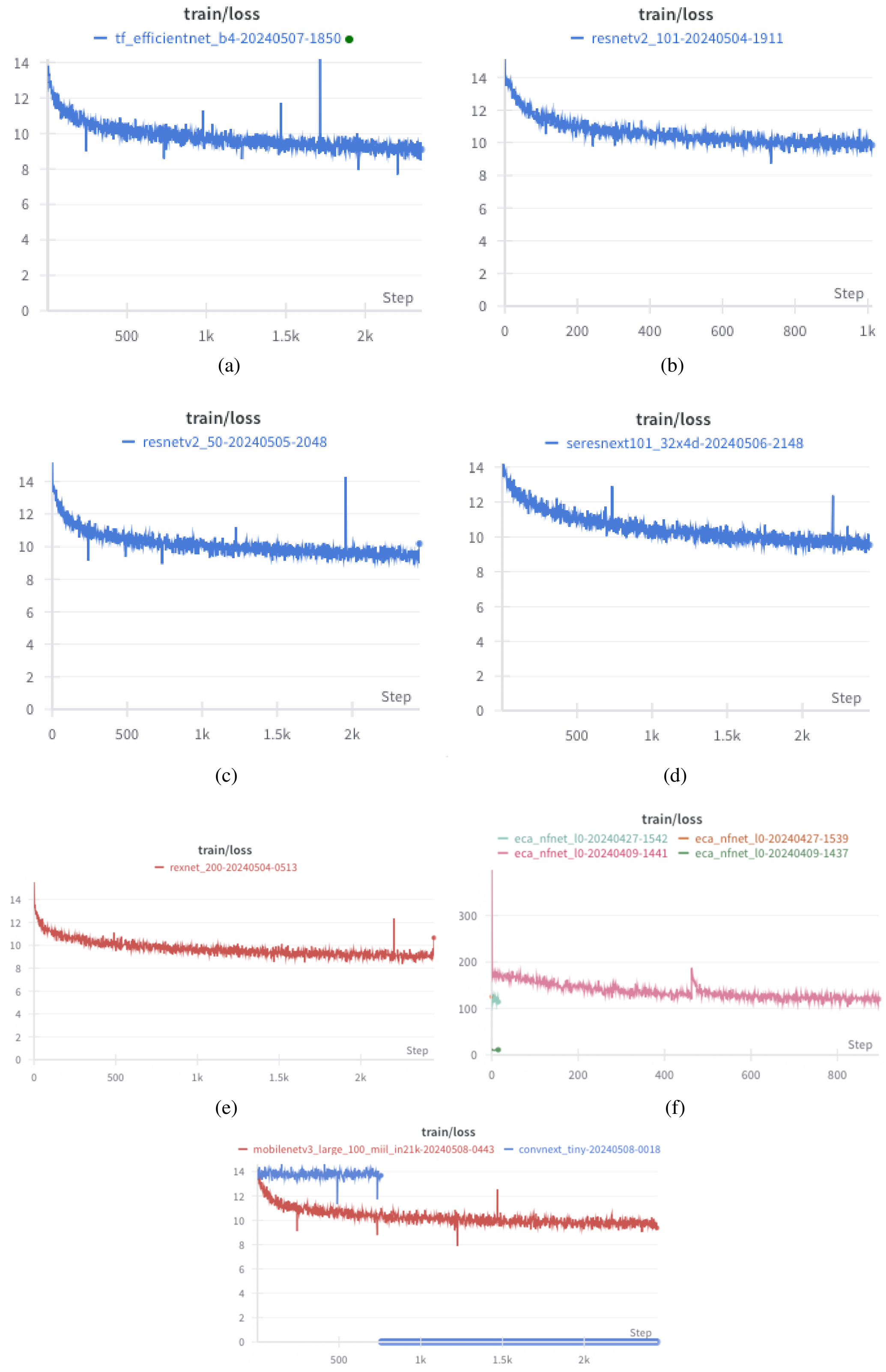
5.3 Presentation Feedback

Overall, the presentation was well-received, with people liking how the slides were well organized and easy to follow. However, there were suggestions to cut down on the technical jargon and make the slides a bit less crowded with text, which could help maintain audience engagement throughout the presentation. Presenters were also advised to also practice more so as to improve their voice modulation, ability to answer questions and interact with the audience, making the presentation more dynamic and engaging.

It was noted that while the Q&A sessions were handled well, they could still be improved. Using more demos or things to visualize the current progress could also make the presentations more interesting and easier to understand as well.

5.4 Technical Feedback

The technical aspects of the presentation received mixed feedback. While the experimental models were well received, many thought that more evidence was needed to prove the model's effectiveness. Reviewers suggested doing quick tests on the models in their early stages to see if they work as expected. This can help everyone understand the direction of the project better and feel more confident about the results. They also said the team should really show how these models can be used in real situations, to prove they actually work well out there in the real world.



(a)

(b)

(c)

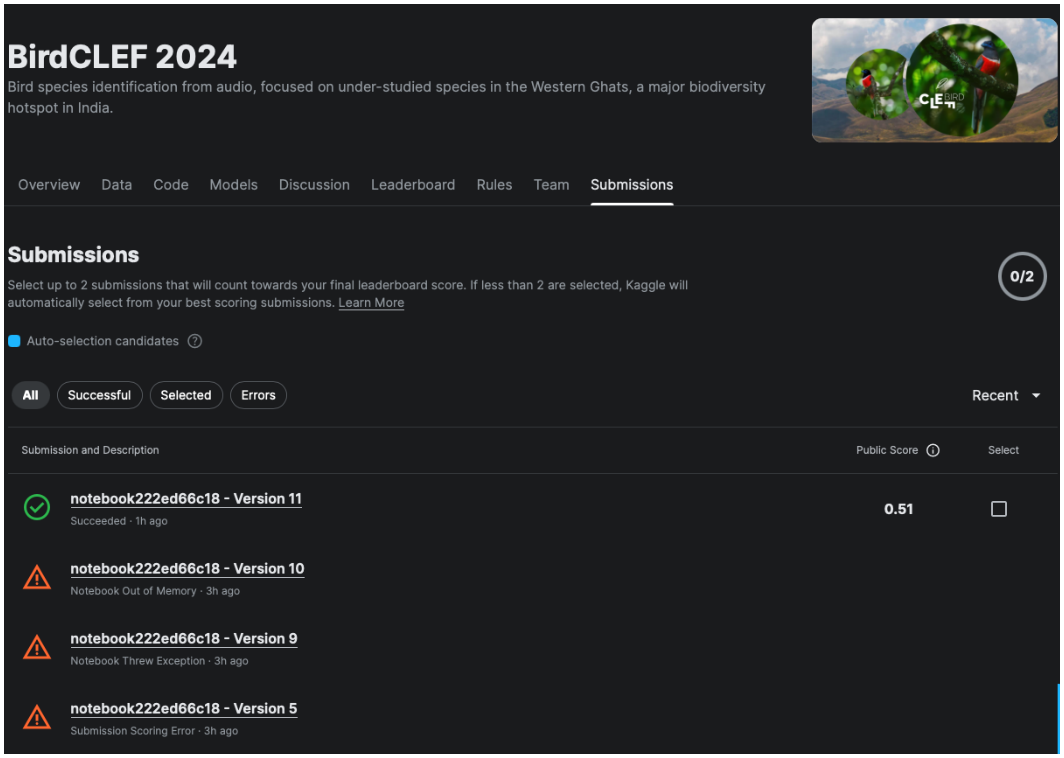
(d)

(e)

(f)

(g)

Figure 2: Train/Loss for (a) EfficientNET model, (b) ResNET V2 101 model, (c) ResNET V2 50 model, (d) SeresNet model, (e) RexNET, (f) EcaNFNet, and (g) MobileNet.



BirdCLEF 2024

Bird species identification from audio, focused on under-studied species in the Western Ghats, a major biodiversity hotspot in India.

Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

Select up to 2 submissions that will count towards your final leaderboard score. If less than 2 are selected, Kaggle will automatically select from your best scoring submissions. Learn More

0/2
Auto-selection candidates
All Successful Selected Errors
Recent

Submission and Description	Public Score	Select
notebook222ed66c18 - Version 11 Succeeded · 1h ago	0.51	□
notebook222ed66c18 - Version 10 Notebook Out of Memory · 3h ago		
notebook222ed66c18 - Version 9 Notebook Threw Exception · 3h ago		
notebook222ed66c18 - Version 5 Submission Scoring Error · 3h ago		

Figure 3: Notebook on Kaggle with working submission (Version 11)